DIRECTIONAL CODING OF AUDIO USING A CIRCULAR MICROPHONE ARRAY

Anastasios Alexandridis*† Anthony Griffin* Athanasios Mouchtaris*‡

* FORTH-ICS, Heraklion, Crete, Greece, GR-70013
† University of Crete, Department of Computer Science, Heraklion, Crete, Greece, GR-71409

ABSTRACT

We propose a real-time method for coding an acoustic environment based on estimating the Direction-of-Arrival (DOA) and reproducing it using an arbitrary loudspeaker configuration or headphones. We encode the sound field with the use of one audio signal and side-information. The audio signal can be further encoded with an MP3 coder to reduce the bitrate. We investigate how such coding can affect the spatial impression and sound quality of spatial audio reproduction. Also, we propose a lossless efficient compression scheme for the side-information. Our method is compared with other recently proposed microphone array based methods for directional coding. Listening tests confirm the effectiveness of our method in achieving excellent reconstruction of the sound field while maintaining the sound quality at high levels.

Index Terms— microphone arrays, spatial audio, beamforming

1. INTRODUCTION

Spatial audio systems aim to reproduce a recorded acoustic environment by preserving the spatial information (e.g., [1, 2, 3, 4]). Such systems have applications in the entertainment sector, enabling users to watch movies that feature surround sound or play computer games providing a more immersive gaming experience, etc. In teleconferencing they can facilitate a more natural way of communication.

In this paper we propose a real-time method for coding a sound field at a low bitrate using microphone arrays and beamforming. Reproduction is possible using an arbitrary loudspeaker configuration or headphones. The sound field is encoded using one audio signal and side-information. We consider microphone arrays—particularly circular arrays—for spatial audio as they are already used in several applications, such as teleconferencing and providing noise-robust speech capture.

Techniques for coding and reproducing spatial audio, when recording a sound scene, have already been proposed. Directional Audio Coding (DirAC) [5] is based on B-format signals and encodes a sound field using one or more signals along with Direction-of-Arrival (DOA) and diffuseness estimates for each time-frequency element. Versions of DirAC that are based on microphone arrays have also been proposed [6, 7]. In [6] differential microphone array techniques are employed to convert the microphone array signals to B-format. However, a bias in the B-format approximation—as illustrated in [8]—leads to biased DOA and diffuseness estimates that can degrade the spatial impression of the result. The authors of [7] utilize array processing techniques to infer the DOA and diffuseness estimates while the reproduction side remains the same as in [5]. Time-frequency array processing is also used in [9] for binaural reproduction.

The aforementioned methods try to encode the sound field in terms of DOA (and diffuseness in the case of DirAC) estimates for each individual time-frequency element, which requires strong W-disjoint orthogonality (WDO) [10] conditions. WDO assumes that there is only one active source in each time-frequency element, which is not the case when multiple sources are active simultaneously. Moreover these methods suffer from spatial aliasing above a certain spatial aliasing cutoff frequency which causes erroneous estimates and can degrade the quality of the reconstructed sound field. Our method tries to overcome these problems by employing a per time frame DOA estimation for multiple simultaneous sources (for details see [11, 12, 13]). Based on the estimated DOAs, spatial filtering with a fixed superdirective beamformer separates the source signals that come from different directions. The signals are downmixed into one audio signal that can be encoded with any compression method (e.g., MP3). Each source signal is reproduced according to its estimated DOA. While the source separation part can create musical distortions in the separated signals, all signals are played back together—since our goal is to recreate the overall sound field—which eliminates the musical noise. This is an important result of our work validated by listening tests.

2. PROPOSED METHOD

Our proposed method is divided into the encoding and the reproduction stage. Both stages are real-time, with the encoding stage consuming approximately 50% of the available processing time—including the DOA estimation and coding of the sound field—on a standard PC (Intel 2.53 GHz Core i5, 4 GB RAM). The reproduction stage can also be implemented in real-time since its main operation is amplitude panning (or HRTF filtering for binaural reproduction).

In an anechoic environment where \( P \) active sources are in the far-field, the signal recorded at the \( m \)th microphone of a microphone array with \( M \) sensors is the sum of the attenuated and delayed versions of the individual source signals according to their direction. Note that although the model is simplified, the experiments presented in this paper are performed using signals recorded in reverberant environments. The microphone array signals are transformed into the Short-Time Fourier Transform (STFT) domain. To estimate the number of active sources and their DOAs, we utilize the method of [11, 12, 13], which is capable of estimating the DOAs in real-time and with high accuracy in reverberant environments for multiple simultaneous sources. The method outputs the estimated number of sources \( \hat{P}_k \) and a vector with the estimated DOAs for each source (with 1st resolution) \( \theta_k = [\hat{\theta}_1 \ldots \hat{\theta}_k] \) per time frame \( k \).

The source signals are then separated using a fixed superdirective beamformer. The beamforming process employs \( \hat{P}_k \) concurrent beamformers each of them steering its beam to one of the directions \( \theta_k \), resulting in the beamformed signals \( B_{\omega k}(\theta, \omega) \), \( s = 1, \cdots, \hat{P}_k \), with \( \omega \) being the frequency index. The beamformer filter coefficients are calculated by maximizing the array gain [14]:

\[
W(\omega, \theta_s) = \frac{\Gamma^{-1}(\omega) \hat{d}(\omega, \theta_s)}{\hat{\tau}^{H}(\omega, \theta_s) \Gamma^{-1}(\omega) \hat{d}(\omega, \theta_s)}
\]

This work is funded by the Marie Curie IAPP “AVID MODE” grant within the 7th European Commission Framework Programme.
where \( w(\omega, \theta_s) \) is the \( M \times 1 \) vector of complex filter coefficients, \( \theta_s \) is the beamformer’s steering direction, \( d(\omega, \theta_s) \) is the steering vector of the array, \( \Gamma(\omega) \) is the \( M \times M \) noise coherence matrix (assumed diffuse), and \((\cdot)^H\) is the Hermitian transpose operation. Fixed beamformers are signal-independent, so they are computationally efficient to implement, facilitating their use in real-time systems, since the filter coefficients for all directions can be estimated offline.

Next, a post-filter is applied to the beamformer output to enhance the source signals. The post-filter constructs \( \hat{P}_s \) binary masks. The mask for the \( s \)th source is given by [15]:

\[
U_s(k, \omega) = \begin{cases} 
1, & \text{if } s = \arg \max_p |B_p(k, \omega)|^2, \quad p = 1, \cdots, P_k \\
0, & \text{otherwise} \end{cases}
\]

Equation (2) implies that for each frequency element only the corresponding element of the source with the highest energy is kept, while the others are set to zero. Thus, the masks are orthogonal. The beamformer outputs are multiplied by their corresponding mask to yield the estimated source signals \( \hat{S}_s(k, \omega), s = 1, \cdots, P_k \).

In this scheme, the number of required bits does not increase linearly with the number of sources. On the contrary, for each next source we need less bits than the previous one. It is computationally efficient, since the main operations are simple OR and NOR operations. The resulted bitstream is further compressed with Golomb entropy coding [17] applied on the run-lengths of ones and zeros.

4. RESULTS

We conducted listening tests on real and simulated microphone array recordings for both loudspeaker and binaural reproduction. We used a uniform circular microphone array with \( M = 8 \) microphones and a radius \( r = 0.05 \) m. The sampling frequency was 44.1 kHz. For loudspeaker reproduction, we used a circular configuration (radius \( 1 \) m) of \( L = 8 \) uniformly spaced loudspeakers (Genelec 8050) and for binaural reproduction we used high-quality headphones (Sennheiser HD650). The coordinate system used for reproduction places the 0\(^{th}\) front in front of the listener, increasing clockwise. The recorded signals were processed using frames of 2048 samples with 50\% overlap, windowed with a von Hann window. The FFT size was 4096.

4.1. Simulated recordings (modelling performance)

We used the Image-Source Method [18] to produce simulated recordings in a reverberant room of dimensions \( 6 \times 6 \times 4 \) m. The walls were characterized by a uniform reflection coefficient of 0.5 and the reverberation time was \( T_{90} = 250 \) ms. The recordings used were a 10-second rock music recording with one male singer at 0\(^{th}\) and 4 instruments at 45\(^{th}\), 90\(^{th}\), 270\(^{th}\), and 315\(^{th}\), which is publicly available from the band “Nine Inch Nails”; a 15-second classical music recording with 6 sources at 30\(^{th}\), 90\(^{th}\), 150\(^{th}\), 210\(^{th}\), 330\(^{th}\), and 270\(^{th}\) from [19]; and a 16-second recording with two speakers, one male and one female, starting from 0\(^{th}\) and walking the entire circle at opposite directions. The recordings included impulsive and non-impulsive sounds. Each source was recorded on a separate track and each track was filtered with the estimated Room Impulse Response from its corresponding direction and then added together to form the array recordings.

The listening tests were based on the ITU-R BS.1116 methodology [20]. Ten volunteers participated in each test (authors not included). For the loudspeaker listening test, each track was positioned at its corresponding direction using VBAP (or by filtering it with the corresponding HRTF for the headphone listening test) to create the reference signals. The low-pass filtered (4 kHz cutoff frequency) reference recording served as quality anchor, while the signal at an arbitrary microphone played back from all loudspeakers (or equally from both left and right channels for the headphone listening test) was used as a spatial anchor. For HRTF filtering, we used the

1The test samples for our method are available at [http://www.ics.forth.gr/~mouchtar/lcasspl3_coding.html](http://www.ics.forth.gr/~mouchtar/lcasspl3_coding.html)
A comparative listening test was conducted with real microphone array recordings. The room dimensions and microphone array specificiations were the same as in Section 4.1. We used an array of Shure SM93 omnidirectional microphones and a TASCAM US2000 USB sound card with 8 channels. The recorded test samples were: a 10-second rock music recording with one male singer at 0° and 4 instruments at 45°, 90°, 270°, and 315°; a 15-second classical music recording with 4 sources at 0°, 45°, 90°, and 270°; and a 10-second recording with two male speakers, one stationary at 240° and one moving clockwise from approximately 320° to 50°. Each source signal was reproduced by a loudspeaker (Genelec 8050) located at the corresponding direction at 1.5 m distance. The sound signals were reproduced simultaneously and captured from the microphone database of [21]. The subjects (sitting at the “sweet spot” for the loudspeaker test) were asked to compare sample recordings against the reference, using a 5-scale grading. Each test was conducted in two separate sessions: spatial impression and sound quality grading.

Our proposed method with two different beamformer cutoff frequencies, namely, \( B = 4 \text{ kHz} \), and \( B = f_s/2 \) (i.e., no diffuse sound) was tested against the microphone array-based methods of [9] and [7]. The extension of [9] for loudspeaker reproduction is straightforward by applying VBAP at each frequency element. The DOA estimation method of [7] is based on the linear array geometry, so we used the localization procedure of [9], combining it with the diffuseness and synthesis method of [7].

The mean scores and 95% confidence intervals for the spatial impression and sound quality sessions for loudspeaker and binaural reproduction are depicted in Figures 1 and 2. An Analysis of Variance (ANOVA) indicates that for both loudspeaker and binaural reproduction a statistical difference between the methods exists in the spatial impression and quality ratings with \( p \)-values < 0.01. Multiple comparison tests using Tukey’s least significant difference at 90% confidence were performed on the ANOVA results to indicate which methods are significantly different. The methods with statistically insignificant differences have been grouped in gray shading.

For both types of reproduction, the best results are achieved with our proposed method when \( B = f_s/2 \) (i.e., no diffuse). With decreasing beamformer cutoff frequency, the spatial impression degrades since directional information is coded only for a limited frequency range. In both versions of our method, the full frequency spectrum is reproduced either from a specific direction or from all loudspeakers (for the diffuse part), so \( B \) does not have a severe impact on the sound quality. Our method, both with \( B \) set to \( f_s/2 \) and 4 kHz receives a better grading than the other methods.

### 4.2. Real recordings (modelling performance)

A comparative listening test was conducted with real microphone array recordings. The room dimensions and microphone array specificiations were the same as in Section 4.1. We used an array of Shure SM93 omnidirectional microphones and a TASCAM US2000 USB sound card with 8 channels. The recorded test samples were: a 10-second rock music recording with one male singer at 0° and 4 instruments at 45°, 90°, 270°, and 315°; a 15-second classical music recording with 4 sources at 0°, 45°, 90°, and 270°; and a 10-second recording with two male speakers, one stationary at 240° and one moving clockwise from approximately 320° to 50°. Each source signal was reproduced by a loudspeaker (Genelec 8050) located at the corresponding direction at 1.5 m distance. The sound signals were reproduced simultaneously and captured from the microphone...
array. The music recordings were obtained from the same sources as in the simulated case. Since a reference recording was not available for this experiment, we employed a preference test (forced choice).

4.3. Simulated recordings (modelling + coding performance)

To investigate how encoding the downmix audio signal with an MP3 encoder affects the spatial audio reproduction, we conducted a listening test with simulated recordings following the same procedure as in Section 4.1. Our proposed method with \( B = f_s/2 \) and \( B = 4 \text{ kHz} \) and the methods of [9] and [7] were included in pairs and the listeners indicated their preference according to the spatial impression and sound quality in two different sessions. The listening test results for all recordings (Table 1) show a clear preference of our method both in spatial impression and quality.

\[
\begin{array}{ccc}
\text{Proposed Huffman} & \text{Proposed Huffman} \\
\text{Rock music} & 140.57 \text{ kbps} & 166.89 \text{ kbps} \\
\text{Classical music} & 128.57 \text{ kbps} & 139.20 \text{ kbps} \\
\text{Speech} & 79.33 \text{ kbps} & 89.10 \text{ kbps} \\
\end{array}
\]

Table 2: Bitrates of the side-information

\[
B = \frac{f_s}{2} \quad B = 4 \text{ kHz}
\]

\[
\begin{array}{cccc}
\text{Proposed Huffman} & \text{Proposed Huffman} \\
\text{Rock music} & 140.57 \text{ kbps} & 166.89 \text{ kbps} \\
\text{Classical music} & 128.57 \text{ kbps} & 139.20 \text{ kbps} \\
\text{Speech} & 79.33 \text{ kbps} & 89.10 \text{ kbps} \\
\end{array}
\]

Fig. 3: Listening test results with MP3 coding at various bitrates for loudspeaker reproduction

Fig. 4: Listening test results with MP3 coding at various bitrates for binaural reproduction

4.3. Simulated recordings (modelling + coding performance)

To investigate how encoding the downmix audio signal with an MP3 encoder affects the spatial audio reproduction, we conducted a listening test with simulated recordings following the same procedure as in Section 4.1. Our proposed method with \( B = f_s/2 \) and \( B = 4 \text{ kHz} \) and the methods of [9] and [7] were included in pairs and the listeners indicated their preference according to the spatial impression and sound quality in two different sessions. The reference and anchor signals were the same as in Section 4.1. We also encoded the side-information using the proposed compression scheme (Section 3). The achieved bitrates for the side-information (with 1° angle resolution for the DOAs) are shown in Table 2. The Golomb parameter \( k \) was set to 2. The bitrates using the Huffman coding on the DOAs are included for comparison. Note that given an angle resolution of 1° and a 4096-point FFT, the required bitrate for the side-information with no coding is approximately 790 kbps for \( B = f_s/2 \) which is comparable to the bitrate of an uncompressed audio signal. The bitrates in Table 2 are different for each recording, since the compression depends on the number of sources and the energy contribution of each source. In the classical music no more than 4 sources are simultaneously active, which explains the smaller bitrate compared to the rock music recording which contains 5 simultaneously active sources.

The mean scores and 95% confidence intervals are shown in Figures 3 and 4. A statistical difference exists both in the spatial impression and sound quality ratings for both reproduction types, based on the ANOVA, with \( p \)-values < 0.01. To indicate which groups are significantly different, we performed multiple comparison tests using Tukey’s least significant difference at 90% confidence. The groups with statistically insignificant differences are denoted with the same symbol at the upper part of Figures 3 and 4. The groups with significantly different ratings are shown with different symbols. It can be observed that 64 kbps achieves the same results as the modelled uncompressed recording both in spatial impression and quality for both \( B = f_s/2 \) and \( B = 4 \text{ kHz} \). Noticeable degradation is evident at 32 kbps. The sound quality degradation is more evident in binaural reproduction, since high-quality headphones allow the listeners to notice more easily small quality impairments caused by MP3 coding. In total, our method can utilize a 64 kbps audio signal plus the bitrate for the side-information to encode the sound field without noticeable degradation in the overall quality caused by the coding procedure.

5. CONCLUSIONS

In this paper a real-time method for encoding a sound field using a circular microphone array was proposed. The sound field is encoded using one audio signal and side-information. An efficient compression scheme for the side-information was also proposed. We investigated how coding the audio signal with MP3 affects the spatial audio reproduction through listening tests and found that coding at 64 kbps results in unnoticeable changes compared with the modelled uncompressed case for the same beamformer cutoff frequency. Comparative listening tests with other array-based methods reveal the effectiveness of our method for loudspeaker and binaural reproduction.
6. REFERENCES


